DOCUMENT RESUME

ED 221 562 TM 820 584

TITLE Local Test Development Manual. Basic Skills Improvement Policy Implementation Guide #5.
INSTITUTION Massachusetts State Dept. of Education, Boston. Bureau of Research and Assessment.
PUB DATE Nov 81
NOTE 112p.

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Achievement Tests; *Basic Skills; Educational Policy; Elementary Secondary Education; Item Analysis; Policy Formation; State Boards of Education; State Programs; Statewide Planning; *Test Construction; Testing Programs; Test Items; Test Reliability
IDENTIFIERS *Basic Skills Improvement Policy (Massachusetts); *Test Manuals

ABSTRACT

This manual is presented to assist local districts in complying with the requirements of the Massachusetts Policy on Basic Skills Improvement, whose principal goal is to help students master basic skills prior to high school graduation. The policy requires schools to evaluate student achievement in the basic skills at least once during the early elementary, later elementary and secondary levels. This manual is designed to provide an overview of the test development process for those districts that plan to develop their own achievement tests at one or more of these levels. The standards and procedures adopted in this manual are based on criteria established by the Massachusetts Department of Education in the areas of content, bias, item quality, and technical quality for the approval of secondary level basic skills tests. These criteria appear in the form of checklists in appendices A-D of this manual. The sections of this manual include: deciding whether to develop a test; identifying basic skills objectives; obtaining items; reviewing items; field testing; constructing test forms; obtaining statistics for test forms; writing a technical manual; obtaining state approval; and a bibliography. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

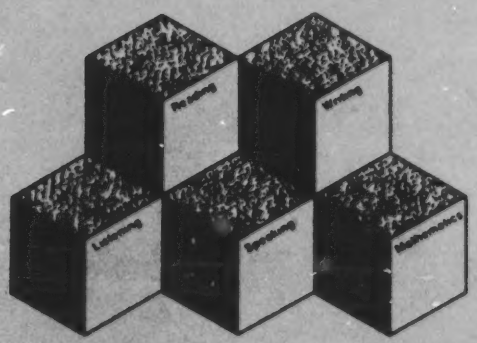
Basic Skills Improvement Policy
Implementation Guide #5

ED221562

LOCAL TEST DEVELOPMENT MANUAL

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C. D. Bell
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."



TM 820 581

Massachusetts Department of Education
November 1981

Massachusetts Board of Education

Ann H. McHugh, Chairperson
James L. Green, Vice Chairperson

John W. Bond
Millie Clements
Joseph P. Crimmins
James R. Grande
Howard A. Greis
Charles T. Grigsby
Mary Ann Hardenbergh
Armando Martinez
Edwin M. Rossman
Donald R. Walker

Michael J. Daly, Acting Commissioner of Education

Bureau of Research and Assessment

Allan S. Hartman, Director
Leslie S. May
Matthew H. Towle
Dorothy Earle
Richards S. Williams
Betty Hancock

The Massachusetts Department of Education insures equal employment/educational opportunities/and affirmative action, regardless of race, color, creed, national origin, or sex, in compliance with Title IX, or handicap, in compliance with Section 504.

Publication #12567, approved by John Manton, State Purchasing Agent

BASIC SKILLS IMPROVEMENT POLICY

IMPLEMENTATION GUIDE #5

LOCAL TEST DEVELOPMENT

MANUAL

Massachusetts Department of Education

November 1981

Copyright © 1981 Massachusetts Department of Education

This publication is restricted for use in the Massachusetts schools for the specific purpose of implementing the Massachusetts Basic Skills Improvement Policy and Regulations and may be duplicated by Massachusetts schools for that purpose. Circulation, distribution, and duplication for use for other purposes requires separate permission from the Massachusetts Department of Education.

TABLE OF CONTENTS

I.	INTRODUCTION	1
II.	DECIDING WHETHER TO DEVELOP A TEST	4
III.	IDENTIFYING BASIC SKILLS OBJECTIVES	7
IV.	OBTAINING ITEMS	9
V.	REVIEWING ITEMS	14
	Validity	14
	Quality	16
	Bias	17
	Readability	17
	Pilot Testing	18
VI.	FIELD TESTING	21
	Conducting the Field Test	21
	Item Analysis	22
	Item Difficulty	24
VII.	CONSTRUCTING TEST FORMS	26
	Number of Forms	26
	Number of Items Per Form	27
	Difficulty	28
	Format	29
	Directions	30
	Practice Questions	30
	Bias Review	30
	Test Administrator's Manual	31
VIII.	OBTAINING STATISTICS FOR TEST FORMS	33
	Test Difficulty	34
	Standard Error of Measurement	36
	Internal Consistency Reliability	39
	Test-Retest Reliability	39
	Parallel Form Reliability	41
	Consistency of Mastery Decisions	43
IX.	WRITING A TECHNICAL MANUAL	46
X.	OBTAINING STATE APPROVAL OF LOCALLY DEVELOPED BASIC SKILLS TESTS	48
	Elementary Level	48
	Secondary Level	48

XI. BIBLIOGRAPHY

General Test Development References	51
Sources of Objectives	57
Sources of Test Items	60
Massachusetts Department of Education Publications	62

Appendix A:	Content Match to Reading Objectives	63
Appendix B:	Bias Review Criteria	69
Appendix C:	Guidelines for Reviewing Item Quality	77
Appendix D:	Technical Review Criteria	79
Appendix E:	Readability Formulas	83
Appendix F:	Standard Error of Measurement	87
Appendix G:	Internal Consistency Reliability Coefficient	91
Appendix H:	Pearson Product-Moment Correlation	95
Appendix I:	Sample Technical Manual	101

I. INTRODUCTION

The purpose of this manual is to assist local districts in complying with the requirements of the Massachusetts Policy on Basic Skills Improvement, whose principal goal is to help students master basic skills prior to high school graduation. To accomplish this goal, the policy requires schools to evaluate student achievement in the basic skills at least once during the early elementary, later elementary and secondary levels. At the early and later elementary levels, districts may choose the evaluation instrument they wish to use. At the secondary level, however, instruments must be approved by the Department of Education. This manual is designed to provide an overview of the test development process for those districts that plan to develop their own achievement tests at one or more of these levels.

The standards and procedures one uses when developing a test depend on how the test is to be used. Obviously, a routine classroom spelling test would not require the kind of intense technical scrutiny as is required of a college entrance examination. The standards and procedures adopted in this manual are based on criteria established by the Department of Education in the areas of (1) content, (2) bias, (3) item quality, and (4) technical quality for the approval of secondary level basic skills tests. These criteria appear in the form of checklists in Appendices A-D of this manual and in the Basic Skills Improvement Policy Implementation Guide #1: Assessing Basic Skills Achievement (Revised Edition), 1980.

While the user of this manual may or may not be developing a test for use in a Basic Skills Improvement Plan, these review checklists provide a convenient set of standards against which any test may be judged, including tests for areas outside of the basic skills. By meeting these standards, one can be assured that the instruments developed are technically sound. If, by chance, the test being developed is for use in a Secondary Level Basic Skills Improvement Plan, it must meet the criteria outlined in these checklists. Careful use of this manual will help to insure that it does.

Besides adopting standards from the Massachusetts Policy on Basic Skills Improvement, there are other ways in which the scope of this manual has been limited. The content has been confined, for instance, to the construction of criterion-referenced* tests, i.e., tests which are designed to measure the mastery of very specific objectives. The collection or interpretation of normative information about test performance has not been addressed. For the most part, discussion and examples have also been limited to tests comprised of multiple choice items. This is not necessarily an endorsement of multiple choice items. They are, however, the most frequently used item type. In reality, almost everything said about multiple choice tests is easily generalizable to tests comprised of other types of items, including items for writing samples. Also omitted from this manual is a discussion of issues related to the scoring of tests or the setting of standards. The latter topic is covered by

---

*The term "criterion-referenced" is used here in its most general sense. Tests developed for the Massachusetts Policy on Basic Skills Improvement might more appropriately be called "objective referenced" or "domain referenced."

Basic Skills Improvement Policy Implementation Guide #3: Standards-Setting Manual (Revised Edition), 1981.

It should be understood that developing one's own test is more difficult than purchasing a commercially prepared instrument. Before undertaking such a task, a district should first be certain that no existing tests adequately measure its objectives. If no test is found, however, a district should approach the test development process with the assurance that it is not an impossible task. It can be done, and, if the proper steps are followed, a test of very high quality can result. It is the purpose of this manual to guide a district through the steps which are essential elements of the test development process. It may be necessary to seek further assistance in implementing some or all of these steps, but at least the scope and sequence of activities will be clear.

The sections of this manual are arranged in sequential order. Section II suggests some factors which should be considered before undertaking the development of a test. Sections III and IV introduce the development of objectives and obtaining items to measure these objectives. Section V suggests procedures for reviewing and pilot testing items. Section VI is a practical guide to conducting a field test and analyzing field test results. Section VII outlines several factors which should be considered when creating test forms from field tested items. Section VIII discusses certain statistics which must be obtained about the test if it is being submitted for state approval. Section IX suggests how those statistics and other relevant data can be compiled in a technical manual, and Section X reviews the process of obtaining state approval. Section XI is a bibliography of books and materials which may be of use in developing a test.

II. DECIDING WHETHER TO DEVELOP A TEST

The decision to develop an achievement test should not be taken lightly. Test development always takes more time and resources than even the most experienced test developers initially estimate. It is important, therefore, that one explore other alternatives and take careful stock of local resources before committing the district to this endeavor.

There are a great many achievement tests on the market which have been developed by commercial test publishers. These tests are not necessarily any better than tests developed by a local district and are unlikely to satisfactorily measure all of a district's educational objectives. However, they are convenient and, in many cases, will end up being cheaper than tests developed from scratch. Before developing a test, therefore, it would behoove a district to thoroughly screen existing instruments. One way to approach this screening task is to apply the criteria outlined in this manual for locally developed tests to tests which have been commercially prepared. Only if existing tests prove unsatisfactory should the district undertake the development of its own instrument.

If the district decides to proceed with the development of its own test, the very first step is to insure that someone is available to coordinate the effort and that sufficient funds and/or release time are available to compensate teachers for the extensive committee work involved. It is also essential that meetings be held and committees be formed early in the process. Conscripting help after test development is underway is a strategy which is certain to fail. There must be

"grassroots" support and commitment to the project to make it viable.

One way to get a realistic picture of what is involved in developing a test is to talk with someone who has done it. The Bureau of Research and Assessment can provide the names of school districts that have developed tests in conjunction with their Basic Skills Improvement Plans. Another way to realistically assess resources is to thoroughly read this manual and roughly calculate the staff, time, and financial resources which would be required to complete each step. Do not begin developing a test until some tentative estimates of the cost have been computed.

The following checklist may help to guide the district through the decision-making process.

STEPS TO CONSIDER
IN DECIDING WHETHER TO DEVELOP A TEST

- Review existing tests to be sure that none of them adequately measures the district's objectives.
- Read Basic Skills Improvement Policy Implementation Guide #5: Local Test Development Manual to get an overview of what steps are involved in developing a test.
- Contact other districts for estimates of time and costs involved.
- Locate a coordinator for the project who has the time, interest, and the ability to supervise all of the steps required.
- Meet with administrators and teachers whose support and assistance will be needed to complete the project. Proceed only if there is "grassroots" support.
- Use the Local Test Development Manual to develop a management plan which includes timelines and a delegation of responsibility for each task.
- Based on the management plan, estimate the costs and amount of teacher release time needed to complete the project.
- Determine whether the district has the financial resources needed to complete the test before it is needed.

III. IDENTIFYING BASIC SKILLS OBJECTIVES

The first step in test development is to spell out very clearly what one is trying to test by developing or adopting a list of objectives which are judged to be representative of the basic skills area (i.e., reading, mathematics, writing, or listening) being measured. Districts with a formal curriculum may already have objectives which represent agreements about the outcomes of instruction in a content or skill area. Districts which do not have formal written objectives must either write their own objectives or adopt objectives written by others. The latter alternative is one which should seriously be considered since writing objectives from scratch will add significantly to the task of developing a test.

There are numerous sources from which objectives may be obtained. The Massachusetts Department of Education, for instance, has published secondary level basic skills objectives in reading, writing, listening, speaking, and mathemtics. These objectives are contained in Section 40.04 of the Basic Skills Improvement Policy and Regulations, and also in Appendix A of this manual. Note that by September 1, 1983, tests used in conjunction with a district's Secondary Basic Skills Improvement Plan must contain at least one item which measures each of these objectives. Most commercial achievement tests list the objectives which individual items are designed tc measure. These objectives are usually found in an administrator's manual or technical manual. In addition, many states and many commercial publishers have developed item and/or objective banks which are for sale or available for the cost of reproduction. A complete

guide to these sources is contained in A Guide to Item Banking in Education (Portland, OR: Northwest Regional Education Laboratory, 1979) by Michael Hiscox. Several of these sources are also listed in Section XI of this manual. The Bureau of Research and Assessment may also be contacted for additional information.

It is important that consensus be reached among local content/skill area specialists and teachers about the relevance and representativeness of the final list of objectives. Since all objectives must be measured, it is wise to limit the number selected to those which are considered essential.

Generally, objectives for test writing purposes should specify at a minimum what behaviors are expected from whom under what circumstances. Objectives should be as specific as possible without being trivial. The more specific and detailed objectives are at this stage, the easier it will be to write test items which measure them. Samples of objectives are included below in EXAMPLE A.

EXAMPLE A
SAMPLE OBJECTIVES

Given the opportunity to use a dictionary, eighth grade students, through their own writing samples, will demonstrate a clear and consistent purpose in their writing.

Fourth grade students will correctly add and subtract four-digit integers with and without regrouping.

Eighth grade students will correctly predict an outcome implied but not stated in a short non-fiction passage from a sixth grade level reader.

IV. OBTAINING ITEMS

Once one has a complete set of objectives, items must be found which measure attainment of these objectives. Again, there are two choices. A district may write its own items, or it may adopt items written by someone else.

If a district chooses to write its own items, it is best to begin by writing item specifications. Item specifications define the precise format of the item so that items can be generated by different people or at different times and still be essentially parallel in form and content. Item specifications identify (1) the general objective to be measured, (2) the specific objective of the item, (3) the format of the item stem (the part of the item which asks the question), and (4) the type of distractors (alternative answers) to be used. Item specifications also provide a sample item written to the preceding requirements. For example, the specification for an item on evaluative comprehension might look something like Example B. A sample specification for rounding whole numbers is contained in Example C.

EXAMPLE B
SAMPLE ITEM SPECIFICATION
FOR EVALUATIVE COMPREHENSION

GENERAL OBJECTIVE

Students will identify a statement as fact or opinion.

ITEM OBJECTIVE

Given a question requiring the students to identify a fact from among a set of four choices, eighth grade students will select the one statement that is a fact from the story.

PASSAGE

There is an unusual hospital in Japan. If you are sick, you could not go to this hospital. This hospital treats sick fish.

For many years the fish hospital has helped fish of all kinds get well. Common fish costing very little money, like guppies, as well as Japanese carp costing five thousand dollars have been treated there. Before the hospital opened, fish owners who loved their pets had to just let their fish die.

One of the doctors at the hospital, Dr. Katoaka, thinks that most ill fish have problems with their scales, fins, or eyes. Many times medicine can be put into the water in the fish tank and the fish get well. Sometimes, however, operations have to be performed on the fish.

Fish owners in Japan are lucky to have a fish hospital. So are the fish!

STEM

Which of the following is a fact from the story?

DISTRACTORS

(1) An opinion stated in the story
(2) An opinion which could be drawn from the story
(3) A fact remotely related to the story

SAMPLE

Which of the following is a fact from the story?
a. Sick fish have problems with their scales, fins, or eyes.
b. All countries should have fish hospitals.
c. The capital of Japan is Tokyo.
d. Operations have been performed on fish.

EXAMPLE C
SAMPLE ITEM SPECIFICATION
FOR ROUNDING OFF WHOLE NUMBERS

GENERAL OBJECTIVE

Students will round off numbers to a specified place.

ITEM OBJECTIVE

Given a question which requires rounding an integer (between 6 and 99999) to a specified place (nearest ten, hundred, or thousand), eighth grade students will select from four choices the correctly rounded value.

STEM

(1) What is _____ rounded off to the nearest _____?
(2) What would be the result of rounding off _____ to the nearest _____?

DISTRACTORS

(1) Integer rounded down when it should have been rounded up or vice versa.
(2) Integer rounded to the next higher place.
(3) Integer rounded to the next lower place.

SAMPLE

What is 1269 rounded off to the nearest hundred?

a. 1200
b. 1000
c. 1270
d. 1300

Item specifications are especially useful when several people are writing items for a single objective, or when additional forms of the test are to be generated in subsequent years. If objectives are sufficiently detailed, it may be possible to write parallel items directly from the objectives without preparing formal item specifications first. However, if the resources are available, it is well worth the effort to develop item specifications prior to writing the items.

If one chooses to adopt existing items rather than to write them, there are several sources to which one can turn. First of all, any test or portion of a test developed by the Massachusetts Department of Education can be used by local districts as a source of items. Secondly, there are many states which have developed entire banks of items for use by local districts. Some of these are referenced in Section XI of this manual. Other states have basic skills tests like those developed in Massachusetts. Items developed by other states are frequently available for a nominal fee. A third source is commercial item banks. Many commercial test publishers maintain banks of items matched to specific objectives. If their objectives are similar to the district's objectives, publishers will either create a test or supply the district with items for the objectives requested. Other sources of items a district may want to consider are professional item writers and other school districts that have developed their own tests.

Please note that ITEMS CONTAINED IN COPYRIGHTED TEST MATERIALS CANNOT BE USED WITHOUT PERMISSION. A district may, of course, use items from commercial achievement tests as examples of item format, but no item may be used verbatim.

While a final decision regarding the number of items which will measure each objective can wait until the items are actually assembled into test forms (see Section VII), it is important to do some preliminary planning at this stage. Make sure enough items are obtained to allow for the loss of unacceptable ones during the review stage, and make sure enough items are obtained to permit the inclusion of multiple items for important objectives.

V. REVIEWING ITEMS

Once items have been written or selected they should be reviewed by a committee of persons other than those who wrote or selected them. Items should be reviewed for validity, technical quality, readability, and bias. They should also be pilot tested on a small number of students at the grade level where they will be used.

VALIDITY. Content validity refers to how well items measure the attainment of the objectives they were designed to assess. Content validity is generally established by an independent check on the match between items and objectives. The process is necessarily somewhat subjective since the result is essentially common agreement among subject matter specialists that a particular item assesses a particular objective.

Perhaps the easiest approach for establishing content validity is to ask a group of content specialists different from the one which developed the items to match items with objectives. Given separate lists of items and objectives, the judges work independently on the matching task. Next, judges' decisions are summarized by calculating the percentage of judges matching an item to the same objective, as in EXAMPLE D.

EXAMPLE D
ESTABLISHING CONTENT VALIDITY
VIA PERCENT AGREEMENT

Percent Agreement Among Judges
For Objective/Item Pairs

Objective	Test Item #	Percent Agreement
A	40	75%
	3	88%
	21	38%
B	6	100%

Another approach for establishing content validity is for judges (again content specialists) to assign numerical ratings indicating the appropriateness of individual items for assessing the objectives they were designed to measure. A common rating scale should be used, and then the ratings for each item/objective pair can be summarized as a mean or average of the scores assigned by judges. In EXAMPLE E, for instance, six judges rated three items from Objective A and one item from Objective B on a four-point scale. The table shows the average rating for each item. Several references cited in Section XI provide additional information about establishing content validity.

EXAMPLE E
ESTABLISHING CONTENT VALIDITY
VIA RATINGS

Average Ratings For Objective/Item Pairs
(1 = Low, 4 = High)

Objective	Test Item #	Average of Judges' Ratings
A	40	2.00
	3	2.75
	21	3.25
B	6	3.50

QUALITY. There are certain rules which professional item writers follow to reduce ambiguity in their items and to insure that items are technically sound. A copy of these standard rules of item writing are contained in Appendix C. Before test forms are constructed, each item

should be examined against this standard set of criteria and revisions made as necessary. If the test is to be used in conjunction with a district's Secondary Basic Skills Improvement Plan, 95% of the items must meet these criteria.

BIAS. Tests can be biased in two different ways. Individual items may be offensive or unfair to certain groups of students, and groups of items may over or underrepresent the interests or experiences of certain groups of students. Consequently, it is important to examine the test for bias at two different stages in its development. First of all, individual items should be reviewed against bias criteria such as those contained in Basic Skills Improvement Policy Implementation Guide #1: Assessing Basic Skills Achievement (Revised Edition), 1980, and in Appendix B of this manual. Note that if the test is to be used in conjunction with a district's Secondary Basic Skills Improvement Plan, it must meet all of these bias criteria. It is usually more efficient to have each reviewer examine the items for a different type of bias (racial, sex, religious, etc.). Later, when the items are assembled into test forms, each form should also be examined to insure that no systematic cultural, political, racial, historical, religious, sex, or social bias has been introduced by the way in which items were assigned to the forms.

READABILITY. It is a good idea to check the reading level of any items which require a substantial amount of reading by students, including mathematics items. Readability formulas are usually designed to be used with passages which are longer than those included in most

tests and thus might tend to be less reliable in the test context. It is therefore suggested that two different readability formulas, such as the Fry Readability Formula, and the SMOG Grading Formula be used with each passage. Copies of these formulas are contained in Appendix E.

For items intended for primary grade students, readability formulas are less useful. In that case, words contained in the item can be checked for level, and commonality of use, against lists such as the Dolch list of sight words, the Dale-Chall list, and/or the local district's basal reader series graded word list. The district's reading specialist should be consulted regarding which method of assessing readability is most appropriate.

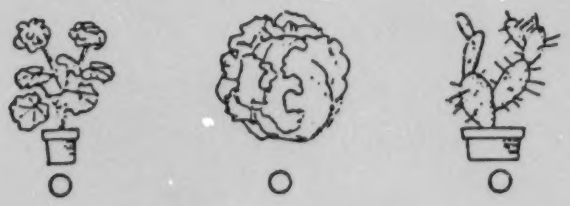
PILOT TESTING. Once items have been reviewed for validity, quality, bias, and readability, they should be tried out on a small sample of children before a formal field test of the items is undertaken. The purpose of this pilot test is to make certain that items are being understood by students taking the test. A child's thought processes are simply not the same as an adult's. Even the most carefully conceived item can prove invalid when submitted to a child's interpretation.

To conduct a pilot test, the test developer should administer newly written items to a small sample of about 10 students who have already mastered the objective. Having students read the items aloud may provide clues to readability level and appropriateness of vocabulary. Children in this sample should be queried about how they arrived at their answers and what specific items they found troublesome. Results of the pilot test will also provide some preliminary feedback regarding item difficulty and time required.

Sources of test item ambiguity which might be identified during a pilot study are:

- confusing directions;
- problems associated with marking the correct answer;
- items which stimulate unusual interpretations;
- visually confusing pictures or graphics;
- items which are subject to personal or egocentric interpretation (especially with primary grade level tests).

An example provided by Haney* provides an illustration of item ambiguity. The question 'Which plant needs the least amount of water? was given orally with respect to the following three illustrations:



One child selected the cabbage as the correct answer and explained his reasoning as follows:

L: Which one needs the least amount of water?

A: "That does."

L: Why is that?

A: "Cause it's a cabbage. Doesn't need as much water. Only when you clean it."

---

*Adapted from Haney, Walt and Laurie Scott, Talking With Children About Tests: A Pilot Study of Test Item Ambiguity, National Consortium on Testing, Staff Circular #7, Cambridge, MA: The Huron Institute, August, 1980.

This child apparently reasoned that the cabbage needed less water than the two potted plants because it was not depicted as a potted cabbage. From this point of view, the logic of the child's reasoning is hard to deny. A head of cabbage once harvested can be kept well and long without water; water is needed "only when you clean it."

Another way of pilot testing newly written, multiple choice items is to provide students with the item stem and allow them to fill in the answer. Not only does this technique provide clues about how students have interpreted the item, but their incorrect answers may later be employed as distractors in a multiple choice format.

Pilot testing is not a required step in the test development process, but one which is highly recommended when reviewing locally developed test items. Items which are obtained from commercial sources or state item banks are likely to have already been pilot tested. If not, they should be pilot tested by the district which adopts them.

VI. FIELD TESTING

CONDUCTING THE FIELD TEST. Once items have been selected and reviewed, it is important to try them out on a representative group of at least 50 students at the grade level where the test is intended to be used. To be representative, this sample should include students from important strata within the district (rural, urban, large schools, small schools, minorities, etc.). This field test will provide data regarding the difficulty of each item and the effectiveness of distractors within each item. Item difficulties will be extremely important when assigning items to test forms, and the number of students who choose each distractor provides additional information about the quality of an item.

For purposes of field testing, items should be grouped by objective. In the field test, it is not necessary for every student to be tested on every objective, but it is important for a given student to receive all of the items for a particular objective. Only by giving all of the items within an objective to the same group of students can the relative difficulties of the items for that objective be accurately determined. With luck several items of similar difficulty will be found which can then be assigned to different forms of the test.

During the field test, therefore, each student will be tested on a subset of the objectives. If there were 30 objectives, for instance, three groups of students might be selected and the items from 10 objectives might be administered to each group. It is important that these groups be representative of students at the grade level for which the test was designed. If classes are "tracked," it is best to choose students of average ability or combine classes of high and low ability.

Better yet would be field testing during "untracked" periods such as homeroom.

Rather than choose three classrooms and administer a different set of items to each, a district might prefer to administer different sets of items to students in the same room. For instance, Student #1 would get the first set; Student #2, the second set; Student #3, the third set; Student #4, the first set again, and so forth. This procedure is called matrix sampling. Matrix sampling helps to insure that a truly representative group of students receives each set of items.

ITEM ANALYSIS. Item analysis can be greatly facilitated if a test scoring service can be found which provides item analysis as part of its scoring package. The most useful item statistic to have on multiple choice items is the number (or percentage) of students who chose each alternative. If tests are hand-scored, this will obviously be a cumbersome statistic to obtain. Someone will have to tally each student's response on each item by hand. The results of such a tally, however, are indispensable in locating faulty items. There are several things one should look for in these data.

1. Items on which more than 25% of the students chose the same wrong answer. One would surely need to investigate why this one distractor was so compelling. EXAMPLE F contains an item of this type.

EXAMPLE F
TEST ITEM WITH FAULTY DISTRACTOR

A sage individual is

a.	touchy	(.02)
b.	old	(.31)
c.	testy	(.02)
*d.	wise	(.65)

Note that nearly one-third of the students chose the same wrong answer. Since the word sage would usually be used to describe someone older than the student, it is not surprising that it connotes age as well as wisdom. If only the less competent students made this mistake, the item is helping to differentiate between them and more competent students. However, if a large proportion of competent students were among the one-third who chose this distractor, the item should probably be revised by replacing distractor b.

2. Items on which one or more distractors were chosen by no one. If one or more distractors are so implausible that they are never chosen, they are not really serving as distractors. If two out of three distractors are never chosen, a multiple choice item becomes essentially a True-False item.

3. Items which over 95% of the students answered correctly. This could mean that the distractors were too obviously incorrect. On the other hand, it could mean that almost all students have mastered that objective. Don't be fooled into thinking that an

item must be faulty if everyone gets it right. Remember that on a criterion-referenced test it is entirely possible that all students have met the criteria for mastery of some objectives. On the other hand, items which everyone answers correctly do not help you to differentiate between masters and non-masters.

4. Items which over 50% of the students answer incorrectly. Again, this pattern of responding could mean one of two things. Either the item is faulty or few students have mastered the objective. On a test of minimum competency, it is unlikely that a large number of students will have failed to meet the minimum standards for mastery. It is important, therefore, that you examine closely any item that few students answer correctly. There may be something about the way in which the correct answer has been phrased which prevents students from recognizing it as being correct.

ITEM DIFFICULTY. Part of an item analysis is the determination of each item's difficulty (p-value). Item difficulty is expressed as the proportion (p) of students who answered the item correctly. In EXAMPLE F, for instance, 65% of the students answered the item correctly. The p-value for that item would therefore be .65.

There is a tendency to think that a test should contain a wide range of item difficulties. In minimum-competency testing, however, it must be expected that the vast majority of students will be at least minimally competent. Thus, the proportion of students who answer the items correctly should be high. The important thing is that an item discriminate those who have mastered an objective from those who have not. If 90% of the

students have mastered a particular objective, then the difficulty for an item which measures that objective should be about .90. If it seems intuitively correct that 90% of the students have mastered the objective and the test developers are satisfied that the item is a valid measure of the objective, do not attempt to make the item harder simply to lower its p-value. On the other hand, if everyone answers an item correctly, that item is not helping to differentiate between competent and incompetent students. Unless it is quite certain that every student has indeed mastered the objective, the district may wish to revise or drop those items with p-values close to 1.00. The only fair criteria to use in such a review are: (1) would someone who has just barely mastered the objective answer the question correctly, and (2) would someone who has not quite mastered the objective answer the question incorrectly. The closer an item comes to measuring the small differences between barely competent and barely incompetent, the more useful that item will be in differentiating between the two groups.

Besides helping to locate items which are technically faulty, item difficulty values can help to screen items for bias. If separate difficulty values are calculated for subgroups of students (e.g., male-female, black-white, rural-urban), items can be located which appear to be especially difficult for one group or another. It is almost certain that some subgroups will perform better than others on the test as a whole, but if one or two items stand out as being significantly more difficult for one subgroup than another, they should probably be dropped.

VII. CONSTRUCTING TEST FORMS

Once field test results have been reviewed and a final set of items is selected for use on the test, the items must be assembled into test forms, accompanied by directions, sample problems, and an administrator's manual. Note that if these forms are to be used in conjunction with a district's Secondary Basic Skills Improvement Plan, they must receive 26 out of 33 points on the technical quality checklist contained in Appendix D. Before constructing test forms you are advised to consult this checklist and to consider the following issues.

NUMBER OF FORMS. Test forms are multiple versions of the same test. Although two forms of a test contain different items, they are presumably measuring the same thing. Because they are measuring the same thing, they should be roughly interchangeable. Since districts are likely to administer a test on more than one occasion, it is convenient to have several forms available. By changing forms of the test each year, for instance, districts can avoid problems associated with students learning, or teachers teaching, the answers to specific items. Parallel forms may also be useful when test administration is staggered within the district, and when students require a makeup test.

The problem with using multiple forms of a test is that no two forms are ever perfectly equivalent. A student who barely passes when given one form of a test might fail when given another. The interform reliability coefficient discussed later provides one measure of how parallel two forms are. On criterion-referenced tests such as those developed for the Massachusetts Policy on Basic Skills Improvement, this lack of perfect

equivalency among forms is only an issue for students who score at or around the mastery cutoff score. For the vast majority of students, slight differences in test forms will not affect the decision of whether or not they are minimally competent. Only students at or around the cutoff can argue that the district's decision may have been different if they had taken a different form of the test. Here again, it is convenient to have multiple forms of the test available. A student who challenges the district's decision can be administered a second form of the test. If a student fails both forms of the test there can be little doubt that (s)he has not mastered the district's basic skills objectives.

The precise number of forms which is developed depends on the resources available to the district. Often it is easier and more efficient to construct several forms at the same time than to add additional forms at a later date. It is usually recommended that at least two forms of the test be developed initially.

NUMBER OF ITEMS PER FORM. The number of test items which are needed to measure each objective depends on how the test is to be used. The more items which measure an objective, the more confident one may be in making decisions about mastery of that objective. Thus, if diagnostic decisions must be made about mastery of a particular objective, it is essential that multiple items measure that objective. In the case of competency testing however, that level of detail is unnecessary. The purpose of the test is to obtain a more general assessment of overall competence. Furthermore, it may not be possible to include multiple items within the context of a mass testing program like that required in basic skills competency testing. If at the secondary level, for instance, a district were to adopt the state's mathematics objectives

and contribute none of their own, the test would have to be a least thirty-eight items long to measure every objective. If two items were included for each objective, the test would be seventy-six items long. At most grade levels, a seventy-six item test is straining reasonable limits for time of administration and student endurance if administered all in one sitting.

Obviously, each objective must be measured by at least one item. Beyond that, the number of additional items which are included depends on the number and importance of the objectives. If there are a few objectives, each may be measured by more than one item. If certain objectives are more important than others, the district may choose to bolster their coverage of those objectives with additional items. In the end, the total number of items on the test should probably not exceed what a borderline minimally competent child in the grade being tested can complete in an hour.

Keep in mind that the initial administration of a basic skills competency test is most often used as a screening device. Students whose competence is in question are likely to be retested or assessed with other instruments to determine the specific nature of their problem. One should not be too concerned, therefore, if the initial screening test does not measure every objective with great precision. More precise measurement is sure to follow.

DIFFICULTY. Parallel forms of a test should be of equal difficulty. Since item difficulties will be available from the field test results, it should be possible to assign items to forms in such a way that the average of the item difficulties in each form is equivalent.

In addition to having the same average difficulty for each form, it is important that items in different forms which measure the same objective also be roughly equivalent in difficulty. Two forms cannot be considered truly parallel if items measuring the same objective have widely discrepant p-values (20 points or more).

FORMAT. When constructing test forms, it is important to pay attention to item layout and other design concerns. The following suggestions should be kept in mind as the test forms are being assembled:

- The number of items per page should be minimized to allow adequate space between items. This is especially important in tests for primary grade children in which three to four items per page may be sufficient.
- Items should not begin on one page and end on another or refer to material on another page unless it is a facing page.
- Multiple choice items within a given section of the test should usually have the same number of alternatives, and response choices should be about the same length.
- Item placement should be consistent; that is, horizontal or vertical order used throughout the test. Horizontal (left to right reading order) is preferable for primary grade children.
- The number of different response formats should be limited to avoid potential confusion in the marking of responses.
- The appearance of the test is important. Care should be taken to select appropriate type style, use legible copies, avoid light-colored inks, and use paper of sufficient weight that print images do not show from the reverse side.
- Figures and drawings should be used carefully. They should be adequate in size and clear in details.
- Subsections or subtests should be clearly labeled so that students can easily find where to start and stop.
- If a separate machine-scorable answer sheet is used, it is best if the number of items on the answer sheet matches the number of items on the test or unused portions of the answer sheet are blocked off. Also make sure that students can find where to begin recording answers for subtests or subsections. In general, it is not wise to use separate answer sheets below the fourth grade.

- Plan to include space for any student-identifier information which will be needed on the answer sheets. This may include student's name or code, grade level, school building, sex, age, homeroom or teacher's name, date of testing, etc.

DIRECTIONS. Directions should be written at a reading level below that which is being tested. Unless an objective of the test is that students read directions for a section independently, directions should be read aloud to students by the test administrator. It is important to insure that test performance is solely a function of the students' responses to the items--not their ability to read the directions (unless that is included in the item specification). To insure consistency across test administrations, explicit instructions should be included in the test administrator's manual regarding how test directions should be given.

Directions should always contain information about the amount of time allowed for completion of the test. They should also contain a policy on guessing. Unless a correction will be made for guessing, it is best to advise students to attempt to answer every question.

PRACTICE QUESTIONS. Every subtest or individually timed subsection should have sample problems which are completed with the assistance of the test administrator. At least one sample problem should be presented for each type of item in a section. This is particularly important when a section contains more than one response format.

BIAS REVIEW. Even though individual items should have been examined for bias before the test forms are constructed, it is important to check

each form for biases resulting from the distribution of items. In this case, it is the configuration of items rather than the items themselves which must be examined. Neither male nor female figures should be underrepresented, nor should any other ethnic, racial, religious, or social group.

TEST ADMINISTRATOR'S MANUAL. To insure consistency across administrations of the test, it is important to develop written guidelines for reference by test administrators. The manual should include the following items:

- any directions to be read aloud to students for the test and its subtests
- a list of materials required by the test administrator and the students
- clear guidelines about the time limits for subsections or subtests, including a policy for handling untimed sections
- directions for explaining practice items
- directions for completion of student-identifier information section
- a policy for answering students' questions about response formats and test content
- for items to be read aloud (e.g., listening test items), guidelines for the number or repetitions allowable of each item
- directions for scoring and summarizing results
- answer keys
- a brief explanation of the purpose of the test, including the use of results

It is best to keep the test administrator's manual brief, including only necessary information.

The following checklist may be useful in reviewing the test forms you have constructed.

ISSUES TO CONSIDER
IN CONSTRUCTING TEST FORMS

- Do test forms contain at least one item per objective?
- Are forms of roughly equal difficulty?
- Were forms designed according to the format considerations on pages 29-30?
- Is the format of the test and answer sheet appropriate for the grade level being tested?
- Do forms contain directions which are suitable for the grade level being tested?
- Do test directions mention time limits and how to handle guessing?
- Does each form contain practice questions?
- Has each form been reviewed for bias?
- Is each form accompanied by a test administrator's manual?

VIII. OBTAINING STATISTICS FOR TEST FORMS

Once test forms have been constructed, it is still necessary to gather information about the finished product: namely, how difficult and how reliable each form is, how comparable the different forms are, and how consistently the test identifies students who are below minimal competency levels. The field test provides information about performance on individual items, but once items have been regrouped into test forms, it is necessary to evaluate performance on the test as a whole.

Although a district may wish to try out its new test right away, this final stage of data collection can usually wait until the first true administration of the test. At that time, the test developers will want to collect information which will permit calculation of test difficulty, standard error of measurement, internal consistency reliability, test-retest reliability, parallel form reliability, and consistency of mastery. These statistics will be needed to meet the technical review criteria contained in Appendix D.

Calculation of the latter three statistics requires that some students be tested twice. Once the test has been administered district-wide, it will be necessary to retest one or more small samples of students (20-30 per sample). Retesting should be done one to three weeks after the initial test and students should be randomly selected for participation in the retest.

One sample of students should receive the same form of the test they took during the district-wide administration. This will permit calculation of a test-retest reliability coefficient (pg. 39) and a

consistency of mastery statistic (pg. 43). The other samples are necessary only if more than one form has been developed. Students in these latter samples should receive a form of the test other than the one they took during the district-wide administration. In other words, students who took Form A should get Form B and students who took Form B should get Form A. This will permit calculation of a parallel form reliability coefficient. Note that in order to calculate the above reliability coefficients, it is necessary to have two scores for each individual student. Be sure that each student's score in these follow-up tests can be matched with that same student's score during the initial test.

TEST DIFFICULTY. Test difficulty is simply the average of the item difficulties, or the average of students' percentage scores converted to a decimal. If the average item difficulty were .80, or, on the average, students answered 80% of the items correctly, the test difficulty would be .80. EXAMPLE G demonstrates these two ways of calculating test difficulty.

EXAMPLE G
TWO WAYS OF
CALCULATING TEST DIFFICULTY

Method 1

Averaging Student Scores

Student	Score
1. Tom	80%
2. Joan	80%
3. Debbie	60%
4. Paul	100%
	80%

Test difficulty = .80

Method 2

Averaging Item Difficulties

Item	p-value
1	1.00
2	.50
3	.75
4	1.00
5	.75
	4.00/5 = .80

Test difficulty = .80

Obviously, there will be a separate test difficulty value for each form of the test. Test difficulty will already have been calculated for each form when trying to equate test forms. However, those estimates were derived from a relatively small sample of students. Furthermore, item difficulties sometimes change as a function of where the items appear in the test. It is therefore wise to recalculate test difficulty after the finished product has been administered district-wide.

As noted above, minimum competency tests should be easy for most students. One should therefore expect the proportion of items answered correctly to be very high. Keep in mind, however, that the test diffi-

culty is an average of individual scores. To get a test difficulty of .80, for instance, three students might have answered 100% of the items correctly, and one student might have answered 20% of the items correctly. Don't let the overall high performance of students in general become a distraction from the main purpose of the test: deciding whether or not the student who answered 20% of the items correctly is minimally competent.

STANDARD ERROR OF MEASUREMENT. Test scores are only estimates of a student's competence or achievement level. Since all estimates contain some error, there is a good chance that a student should have scored a little better or a little worse than (s)he actually did. A statistic called the standard error of measurement provides an estimate of how much error is likely present, and allows one to calculate a range of scores within which a student's true achievement level probably lies. The only students for whom this error is likely to have any serious consequences are students who are near the cutoff score which the district has established for mastery or minimum competence. If, because of measurement error, a student near the cutoff has scored higher than (s)he really should have, that student may be judged competent when (s)he actually is not. Conversely, a student who scores lower than (s)he really should have, may be judged not competent when (s)he actually is. In short, students can easily fall on the wrong side of the cutoff score because of measurement error.

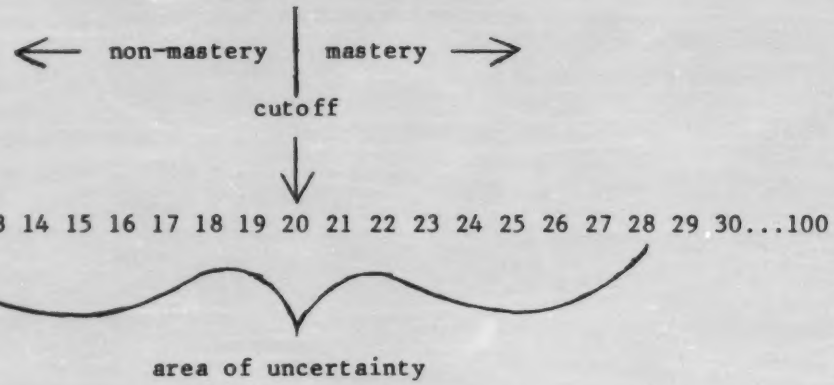
One way to prevent misclassification of students is to identify a range of scores around the cutoff which are "too close to call."

Students within this range might be above the cutoff or below the cutoff, but because of measurement error in the test, it is impossible to tell which. Students within this "range of uncertainty" around the cutoff could then be retested with another form of the test, or arbitrary decisions could be made about where they belong. The width of this range around the cutoff is a function of the standard error of measurement at the cutoff. Appendix F describes how to estimate the standard error of measurement for a given cutoff score. Standard error of measurement is also discussed on pages 36 and 38 of the Basic Skills Improvement Policy Implementation Guide #3: Standards-Setting Manual (Revised Edition), 1981. Once the standard error has been calculated, the width of the range is usually defined as two standard errors above to two standard errors below the cutoff score. The following example (EXAMPLE H) demonstrates the use of standard error in this way.

EXAMPLE H
STANDARD ERROR OF MEASUREMENT

Suppose on a test of 100 items, the cutoff were set at 20. Following the steps in Appendix F we find that the standard error of measurement for a score of 20 on a test of 100 items is 4.

Since the range of uncertainty around the cutoff encompasses the area from two standard errors below to two standard errors above the score, we know that any student with a score between 12 (i.e., 20 - 2(4)) and 28 (i.e., 20 + 2(4)) should be examined more carefully before a final decision is made.



Although standard error of measurement has been discussed only in terms of its implications for standard setting, it can actually be used to establish a "confidence band" around any student's score. In general, 95% of the time a student's true level of achievement will lie within the range from two standard errors above to two standard errors below the score (s)he obtains on a test.

INTERNAL CONSISTENCY RELIABILITY. When one develops a basic skills test, there is an implicit assumption that all of the items on the test are measuring the same basic skill. Math items, for example, should not measure reading ability; reading items should not measure a student's knowledge of history. A total score on the test will have little meaning unless that assumption is true. To be sure that a test is internally consistent (i.e., that test items are measuring the same basic skill), it is important to calculate an internal consistency reliability coefficient. The most common coefficient of this type is the Kuder-Richardson Formula 20, better known as the KR-20. To obtain this coefficient, one needs the item difficulty for each item (see page 24), as well as the score for each student. Complete instructions for calculating the KR-20 are contained in Appendix G.

The value of the coefficient can range from 0 to +1.00. The more internally consistent the test is, the higher the coefficient will be. In general, a KR-20 of .90 or greater is considered acceptable. A coefficient of less than .80 suggests that a major revision of the test is needed. While there are statistical techniques that will indicate which items are lowering the test's internal consistency, it is usually possible to find such items by simply reviewing test scores and item difficulties. Items which are answered correctly by students with low scores but incorrectly by students with high scores should be considered suspect.

TEST-RETEST RELIABILITY. A test provides an estimate of a student's

competence or level of achievement. One of the main questions which must be answered about a test is whether repeated administrations will always give the same estimate. It is only necessary to administer the test twice to determine this type of reliability (see EXAMPLE I). As described in the section on field testing, a small sample of students should be retested and a Pearson product-moment correlation calculated between their scores on the first and second administration. This correlation coefficient is called a test-retest reliability coefficient. Calculation of the Pearson product-moment correlation is described in Appendix H. It is also covered in most elementary statistics textbooks. The value of the coefficient (r) can range from -1.00 to +1.00. If students who scored high the first time also scored high the second time, and students who scored low the first time also scored low the second time, the value of r should be a high positive number. To be considered adequately reliable, the test-retest reliability coefficient should be +.90 or higher. A reliability coefficient of less than +.80 should be regarded as unacceptable. If there were no relationship between the students' first scores and their second scores, the value of r would be zero; if there were an inverse relationship, the value of r would be negative.

EXAMPLE I
TEST-RETEST RELIABILITY

Student	Score On First Administration	Score On Second Administration
Amanda	93	89
Andrew	85	83
Becky	79	81
Bill	82	83

Test-Retest Reliability between first and second administrations

r = .97

(see Appendix H for calculation instructions)

PARALLEL FORM RELIABILITY. When more than one form of a test has been developed, there is more than one way of obtaining an estimate of a student's competence or achievement level. Ideally, different forms should give the same estimate, but a perfect relationship is rarely attained.

To determine how close two forms come to giving the same results, it is necessary to administer both forms to a sample of students and calculate a Pearson product-moment correlation between the two sets. This correlation is called a parallel form reliability coefficient. Calculation of the Pearson product-moment correlation is described in Appendix H. It is also covered in most elementary statistics textbooks. The value of the coefficient (r) can range from -1.00 to +1.00. If students who scored high on the first form also scored high on the second, the value of r should be a high positive number. If there were no relationship between students' scores on the two forms, the value of r would be zero; if there were an inverse relationship, the value of r would be negative.

To be considered parallel, two forms of a test should have a reliability coefficient of +.90 or higher. If the coefficient is below +.80, the two should definitely not be considered parallel. A high reliability coefficient usually means that the two forms are measuring the same thing, but it does not mean that scores on one are equivalent to scores on the other. Test forms can be very different in overall difficulty and still have high parallel-form reliability. Unfortunately, there is no statistically correct method for equating one form to another using difficulty (p) values. In other words, if Form A has a p-value of +.78 and Form B has a p-value of +.88, one cannot equate the two by simply adding a .10 to all of the scores on Form A.

There are ways of equating test forms (e.g., Rasch scaling), but these methods are beyond the scope of this manual and beyond the resources of most districts. It is possible to equate the cutoff scores

on parallel forms, however, simply by choosing a cutoff score on the second form which produces the same percentage of failures as the cutoff on the first. For example, if a cutoff score of 21 items correct on Form A resulted in 10% of the students failing the test, a cutoff score should be chosen on Form B which also results in 10% of the students failing the test. This equating is best done with data from the sample of students who took both forms for purposes of calculating parallel form reliability.

CONSISTENCY OF MASTERY DECISIONS. Another statistic which helps to identify the amount of error associated with mastery decisions is the percentage of agreement between two administrations of the same test. In other words, if a test identifies a student as a non-master, would a second administration of the same test lead to the same conclusion? This information can easily be obtained from the data which is collected for calculating test-retest reliability. Each student in the test-retest sample will have two test scores from the same form of the test. To obtain the consistency of mastery statistic simply count the percentage of students who fell on the same side of the cutoff score the second time as they did on the first. EXAMPLE J demonstrates this procedure.

EXAMPLE J
CONSISTENCY OF MASTERY

If a sample of 30 students were tested twice on the same test, one could record mastery/non-mastery decisions on a table as follows:

		First Administration: Mastery (Above Cutoff)	First Administration: Non-Mastery (Below Cutoff)
Second Administration	Mastery (Above Cutoff)	25	1
	Non-Mastery (Below Cutoff)	0	4

Twenty-five students scored above the cutoff and four students scored below the cutoff both times. One student who scored below the cutoff the first time scored above it the second, i.e., the first time this student was classified as a master and the second time as a non-master. Nevertheless, 97% of the students (29 out of 30) were classified the same way the second time as they were the first. It is this percentage which indicates consistency of mastery decisions.

If the district has not established a cutoff score but wishes to calculate consistency of mastery, select an arbitrary score somewhere close to where the cutoff is likely to be, or select several arbitrary cutoff scores and average the percentage of agreement obtained with each of them. A widely accepted standard for consistency is 90%. If fewer than 90% of the students are classified in the same way on a retest as they were initially, the district should make an effort to reevaluate those students whose scores are near the cutoff because the test is obviously not permitting reliable decisions to be made.

This same approach can be used to determine the amount of error associated with mastery decisions when two different forms of a test are administered. In that case, the percentage of agreement would be the percentage of students who are classified in the same way on one form of a test as they were on another. This information can easily be obtained from the data which is collected for calculating parallel form reliability.

IX. WRITING A TECHNICAL MANUAL.

The technical manual should document all of the preceding steps in test development, especially the results of the field test. The technical manual need not be long, but must contain all of the information which is necessary for an outsider to judge the quality of the test. The Massachusetts Department of Education does not require development of a technical manual for tests used in conjunction with the Basic Skills Improvement Policy. However, the information which is normally contained in a technical manual must be available to the public and must be submitted to the Bureau of Research and Assessment if state approval is being sought for use of the test at the secondary level. It is therefore to the district's advantage if all necessary information is collected in a single document like a technical manual. Simply having a technical manual adds credibility to the test development process.

A typical technical manual would have the following outline.

I. Objectives

II. Item Review

- A. Quality
- B. Content Validity
- C. Bias
- D. Pilot Testing

III. Test Forms

- A. Format
- B. Directions
- C. Sample Items
- D. Bias Review

IV. Field Testing

A. Sampling Plan
B. Item Analysis
C. Item Difficulty

V. Test Statistics

A. Test Difficulty
B. Standard Error of Measurement
C. Internal Consistency
D. Test-Retest Reliability
E. Parallel Form Reliability
F. Consistency of Mastery Decisions

A sample technical manual is contained in Appendix I.

X. OBTAINING STATE APPROVAL OF LOCALLY-DEVELOPED BASIC SKILLS TESTS

ELEMENTARY LEVEL. State approval is not required for basic skills tests at the elementary level. However, districts are still urged to follow the procedures outlined in this manual so that the instrument which is developed can withstand any challenges which arise regarding its technical quality. Furthermore, it is clearly in the district's best interests to produce a high-quality instrument--one which is valid, reliable, and free from bias. Following the procedures outlined here will help to inspire confidence in the scores which are obtained and the decisions which are made on the basis of those scores.

SECONDARY LEVEL. State approval is required for basic skills tests in reading, mathematics and listening at the secondary level.* Procedures for obtaining approval are outlined in Basic Skills Improvement Policy Implementation Guide #1: Assessing Basic Skills Achievement (Revised Edition), 1980. In short, a district must:

(1) Inform the Bureau of Research and Assessment that it wants to use a locally developed instrument.

(2) Review the test according to the content match (Appendix A), bias review (Appendix B), item quality (Appendix C), and technical quality (Appendix D) criteria established by the Department of Education.

(3) Provide 5 copies of the test, answer sheet, directions, and technical information to the Department. Technical information may be provided in any format the district chooses as long as the district addresses relevant portions of the Technical Review Criteria contained in Appendices A-D of this manual.

---

*State approval is not required for writing tests. In writing, only assurances are required.

For listening tests, districts must also provide the Department with the following assurances regarding the testing situation.

a. The testing environment is free of excessive noise, static, echoing, or other distractions that would interfere with students' ability to hear orally presented materials.

b. Oral test materials are presented with clarity and at a volume, pitch and speed that will not interfere with students' ability to hear orally presented materials.

c. A tape recorder is provided that is functioning properly (if such equipment is used).

d. Instructions provided by the test publishers are followed carefully (if material is to be read aloud).

e. Whether students can hear orally presented materials without difficulty is ascertained before beginning the actual testing.

Writing tests do not require state approval. However, districts must provide the Department with the following assurances regarding topics, testing situations, and scoring procedures.

a. All students doing the writing sample are provided an opportunity to use a dictionary.

b. Scoring procedures are used that assess student performance on the secondary level writing objectives set forth in Section 40.04 of the Regulations.

c. Scoring procedures are used that insure reliability of results.

Assurances must also be given that any commercial or locally developed instrument used to assess writing skills has the following characteristics:

a. Each item is clear and concise.

b. The directions are specific.

c. The topic allows for a range of responses.

d. The vocabulary is not too difficult to be understood by an average student.

e. Each item indicates the way the student should proceed.

f. Each item is within the range of experience of all students.

g. Each item is free of offensive sex, cultural, racial, and ethnic content and stereotyping.

The Department will review all information submitted. If the local test is approved, districts will be notified along with formal notification of action on the entire Plan. If the test does not meet the requirement for approval, the district will be notified as soon as possible. Notification will include specific additional requirements to be met for approval.

XI. BIBLIOGRAPHY

A. GENERAL TEST DEVELOPMENT REFERENCES

Ahmann, J.S. and Glock, M.D. Evaluating Pupil Growth: Principles of Tests and Measurements (5th ed.). Boston: Allyn & Bacon, 1975.

- Introductory text--see discussions of educational objectives and measuring complex achievement objectively.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. Standards for Educational and Psychological Tests and Manuals. Washington: American Psychological Association, 1974.

- Three-level standards (essential, very desirable, desirable) for tests, manuals, and reports, research on reliability and validity, use of tests.

Anastasi, A. Psychological Testing (4th ed.). New York: Macmillan, 1976.

- Discussion of principles of testing and test construction, reliability, validity, and item analysis.

Berk, R. (Ed.). Criterion-Referenced Measurement: State of the Art. Baltimore: The Johns Hopkins University Press, 1980.

- Collection of conference papers. Includes articles by Popham on content domain specifications, Berk on item analysis, Hambleton on test score validity and cut-off scores among others.

Bloom, B.S..(Ed.). Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain. New York: Longman, 1977.

- A system for classifying objectives which offers a framework for building evaluation instruments. Also illustrative test items.

Bloom, B.S., et.al. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw Hill, 1971.

o Many examples of objectives, illustrative test items and evaluative techniques for levels of Bloom's taxonomy in various subject areas.

Brown, Frederick G. Principles of Educational and Psychological Testing (2nd ed.). New York: Holt, Rinehart and Winston, 1976.

o General textbook on testing, emphasizing use and evaluation of tests. See chapters on consistency and validity and interpreting test scores.

Cronbach, L.J. Essentials of Psychological Testing (3rd ed.). New York: Harper and Row, 1970.

o Discussion of types of validity. Steps for computing statistics related to reliability.

Ebel, R.L. Essentials of Educational Measurement (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall, 1979.

o See "How to Write Multiple-Choice Test Items" and "How to Improve Test Quality Through Item Analysis."

Gronlund, N.E. Constructing Achievement Tests (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, 1977.

o Practical and readable guidebook for test construction. Includes discussion of determining reliability, validity and conducting an item analysis.

Gronlund, N.E. Measurement and Evaluation in Teaching (3rd ed.). New York: Macmillan, 1976.

o See Chapter 8, "Constructing Objective Test Items: Multiple-Choice Form;" Chapter 2, "Defining Objectives for Evaluation Purposes."

Gronlund, N.E. Preparing Criterion-Referenced Tests for Classroom Instruction. New York: Macmillan, 1973.

Gronlund, N.E. Stating Objectives for Classroom Instruction (2nd ed.). New York: Macmillan, 1978.

- Practical and short booklet designed to teach specification of instructional objectives. Includes example of using Bloom's taxonomy, checklist for evaluating objectives, list of verbs useful for defining objectives.

Guilford, J.P. and Benjamin Fruchter. Fundamental Statistics in Psychology and Education (6th ed.). New York: McGraw Hill, 1977.

- Basic source book for information about test-related statistics. Includes sections on reliability, item analysis, validity.

Hambleton, Ronald K. "Validation of Criterion-Referenced Test Score Interpretations and Standard Setting Methods," Laboratory of Psychometrics and Evaluative Research Report No. 85. Amherst, MA: University of Massachusetts, 1978.

Hannah, Larry S. and Michaelis, John U. A Comprehensive Framework for Instructional Objectives. Reading, MA: Addison-Wesley Publishing Company, 1977.

- Describes a framework for developing instructional objectives based on hierarchies within intellectual processing, skills, and attitudes and values. Many examples of objectives and items.

Harris, C.W., Alkin, M.C. and Popham, W.J. (Eds.). Problems in Criterion-Referenced Measurement. CSE Monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

- Collection of short essays which address practical problems. Topics include selection of objectives, item generation, selection of items, test length.

Hively, W. (Ed.). Domain Referenced Testing. Englewood Cliffs, NJ: Educational Technology, 1974.

- A collection of articles, emphasizing applications of domain referenced testing.

Kryspin, W.J., and Feldhusen, J.F. Developing Classroom Tests: A Guide for Writing and Evaluating Test Items. Minneapolis, MN: Burgess, 1974.

McAshan, H.H. Writing Behavioral Objectives: A New Approach. New York: Harper and Row, 1970.

- Guide to mechanics of writing performance objectives using a goals approach. Includes sections on developing goal statements and identifying evaluation activities.

Mager, Robert F. Measuring Instructional Intent, or Got a Match? Belmont, CA: Fearon Publishers, 1962.

- Mager describes a procedure for matching test items to objectives. Includes practice in locating and interpreting key characteristics of an objective, matching performances and matching conditions.

Mager, Robert F. Preparing Instructional Objectives (2nd ed.). Belmont, CA: Fearon Publishers, 1962.

- By means of a branching format, Mager demonstrates how to define and state teaching objectives and how to describe criteria used to measure success.

Morris, Lynn and Fitz-Gibbon, Carol T. How to Deal with Goals and Objectives. Beverly Hills, CA: Sage Publications, 1978.

- Practical advice on locating and writing objectives. Chapter 4 is a complete discussion of assigning priorities to objectives.

Morris, Lynn and Fitz-Gibbon, Carol T. How to Measure Achievement. Beverly Hills, CA: Sage Publications, 1978.

- Practical guide for selecting a test to match objectives. Includes comprehensive list of resources for test construction, checklist of item construction errors.

Nagel, Thomas S. and Richman, Paul T. Competency-Based Instruction. Columbus, OH: Charles E. Merrill Publishing Company, 1972.

- A branching programmed text, emphasizing development of instructional objectives as basis for mastery learning.

Payne, David A. The Assessment of Learning: Cognitive and Affective. Lexington, MA: D.C. Heath and Co., 1974.

- Text emphasizes relationship of test items to instructional objectives and techniques for writing test items. Chapter 2 discusses specifications of cognitive objectives. Chapter 12 is about item analysis.

Popham, W.J. Criterion-Referenced Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978.

- Discussion of definition of domains, reliability and validity of criterion-referenced measures.

Popham, W.J. Educational Evaluation. Englewood Cliffs, NJ: Prentice-Hall, 1975.

- Chapters on objectives and criterion-referenced measurement.

Popham, W.J. (Ed.). Criterion-Referenced Measurement: An Introduction. Englewood Cliffs, NJ: Educational Technology Publications, 1971.

- A short collection of articles and symposium papers including discussions of technical constructs related to construction of criterion-referenced tests.

Popham, W.J., (Ed.). Evaluation in Education: Current Applications. Berkeley, CA: McCutchan, 1974.

- Includes chapter by Millman on criterion-referenced measurement.

Saunders, Norris M. Classroom Questions: What Kinds? New York: Harper and Row, 1966.

- Samples of questions for varying levels of complexity based on Bloom's taxonomy.

Thorndike, R.L. (Ed.) Educational Measurement (2nd ed.). Washington: American Council on Education, 1971.

- A comprehensive source book, presenting expert opinion on test development topics. See Wesman on item writing, Cronbach on test validation, Kratwohl on definition of objectives, Tinkelman on planning the objective test.

Tuckman, Bruce W. Measuring Educational Outcomes: Fundamentals of Testing. New York: Harcourt, Brace, Jovanovich, Inc., 1975.

- General introductory textbook for teachers. Includes chapters on construction of objectives, developing short-answer items, validity, and reliability.

Womer, F.B. Basic Concepts in Testing. New York: Houghton-Mifflin, 1968.

- Brief and straightforward discussion of basic testing concepts such as error of measurement, interpretation of reliability coefficients, test validity.

B. SOURCES OF OBJECTIVES

Collections and Distributors*

CCL Document Service
1025 W. Johnson Street
Madison, WI 53706

- An ordered list of K-8 math objectives accompanied by 400+ sample test items.

Cambridge Book Company
488 Madison Avenue
New York, NY 10022

- Learning and behavioral objectives at primary and secondary levels.

Commercial-Educational Distributing Services
P.O. Box 8723
Portland, OR 97208

- Course goals developed by Tri-County project.

The Co-op
413 Hills House North
University of Massachusetts
Amherst, MA 01002

- Objectives and items for K-12 in language arts, math and other subjects.

Department of Curriculum and Instruction
Professional Sequence
Pre-Science Secondary Teacher Education
Illinois State University
Normal, IL 61761

- Contact source for further information

---

*Most of the organizations listed will charge a fee for materials. Inclusion is not an endorsement of objectives. School districts should make their own judgements about the appropriateness of objectives. The Department has not reviewed these objectives collections.

Educational Research Associates
Bureau of General and Academic Education
Department of Education
Box 911
Harrisburg, PA 17126

- PRIMES Objectives: math.

Florida Center for (Competency-Based) Teacher Training Materials
College of Education
University of Miami
Miami, FL 33100

- Contact source for further information

IPI Objectives
c/o Dr. Robert Scanlon
1700 Market Street
Philadelphia, PA 19103

- Math, science and reading.

Institute for the Development of Educational Advancement
(IDEA) University of California
Suite 950
1100 Glendon Avenue
Los Angeles, CA 90024

- UNIPAC Objectives in all subject areas.

Institute for Educational Research
1400 W. Maple Avenue
Downers Grove, IL 60515

- Pool of behavioral objectives and test items for K-12 in language arts, math, social science and science. (ERIC: ED 066 492 through 508)

Instructional Objectives Exchange
Box 24095
Los Angeles, CA 90024

- Separate booklets of objectives, including sample test items, for language arts, mathematics, English grammar, reading and other subjects.

Regional Education Lab for the Carolinas and Virginia
613 Vickers
Durham, NC 27701

- Individualized Mathematics System Objectives.

Westinghouse Learning Corporation
Palo Alto, CA

- Behavioral objectives--a guide to individualizing learning. Objectives for primary, intermediate and secondary levels in mathematics, language arts, science and social studies.

- Learning objectives for individualized instruction in social science, science, language arts and mathematics.

Reference Books

Bloom, B.S. (Ed.). Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. New York: Longman, 1977.

- A comprehensive framework for classifying objectives.

French, W. et.al. Behavioral Goals of General Education in High School. New York: Russell Sage Foundation, 1957.

- Comprehensive list of objectives.

Kearney, N.C. Elementary School Objectives. New York: Russell Sage Foundation, 1953.

- Comprehensive list at elementary level in nine curriculum areas and four behavioral patterns (knowledge and understanding, skills and competencies, attitudes and interests, actions).

National Council of Teachers of Mathematics. Evaluation in Mathematics, Twenty-sixth yearbook. Washington: The National Council, 1961.

- Illustrative objectives.

C. SOURCES OF TEST ITEMS

California Department of Education
c/o Dr. William Padia
721 Capitol Mall
Sacramento, CA 95814

- 628 items in reading, 628 items in math, and 501 items in language for grades 4, 6, 7, 9, 11.
- Tests in reading and math for grades 8 and 10.

Delaware Department of Public Instruction
c/o Alice Valdes
P.O. Box 1402
Dover, DE 19901

- 252 items in reading, 543 items in math for various grades.

Fountain Valley Teacher Support System
c/o Richard Zweig Associates
20800 Beach Blvd.
Huntington Beach, CA 92648

- 1559 items in reading and 2674 items in math for various grades K-adult.

Los Angeles County Test Development Center
c/o John Martois
9300 E. Imperial Highway
Downey, CA 90242

- 2000 items in reading, 1900 items in math, and 1700 items in language arts for grades K-adult.

---

*Most of the organizations listed will charge a fee for materials. Inclusion is not an endorsement of items. School districts should make their own judgements about the appropriateness of items. The Department has not reviewed these item banks.

Michigan Department of Education
c/o Ed Roeber
P.O. Box 30008
Lansing, MI 48909

- 1000 items in reading and 3000 items in math for grades 4, 7, and 10.
- Tests available in reading and math for grades 4, 7, and 10.

National Assessment of Educational Progress
c/o Jack Schmidt
Suite 700
1860 Lincoln Street
Denver, CO 80295

- 134 items in reading, 494 items in math, and 63 items in language arts for ages 9, 13, 17, and adult.

New Jersey Department of Education
c/o Dr. William Mathis
225 West State Street
Trenton, NJ 08625

- Tests in reading and math for grades 3, 6, 9 and 11.

Northwest Evaluation Association
c/o LeAnn Gamache
P.O. Box 1357
Tacoma, WA 98401

- 1167 items in reading and 1467 items in math for various grades from K-12.

U.S. Department of Defense
c/o Office of Dependents Schools
2461 Eisenhower Avenue
Alexandria, VA 22331

- Tests in reading, math and language arts in various grades from 3-11.

D. MASSACHUSETTS DEPARTMENT OF EDUCATION PUBLICATIONS

Basic Skills Improvement Policy and Regulations. Boston, MA: Massachusetts Department of Education, January, 1979.

- Text of the basic skills policy and regulations.

Basic Skills Improvement Policy Implementation Guide #1: Assessing Basic Skills Achievement (Revised Edition). Boston, MA: Massachusetts Department of Education, May, 1980.

- District and state responsibilities; assessment issues; assessment at the elementary level; assessment at the secondary level; special needs; reporting results.

Basic Skills Improvement Policy Implementation Guide #2: Writing Assessment Manual (Revised Edition). Boston, MA: Massachusetts Department of Education, March, 1981.

- Instructions for administrators, readers, chief readers, and aides.

Basic Skills Improvement Policy Implementation Guide #3: Standards-Setting Manual (Revised Edition). Boston, MA: Massachusetts Department of Education, March, 1981.

- Item review methods; methods based on judgments of student's mastery status; interpreting and using standards; regulations.

Basic Skills Improvement Policy Implementation Guide #4: Basic Skills Improvement Policy Forms and Procedures. Boston, MA: Massachusetts Department of Education, December, 1979.

- Affirmation of elementary level plans; submission of secondary level plans; suggestions for the development of plans; review of plans.

Appendix A

Massachusetts Secondary Basic Skills Objectives

MASSACHUSETTS DEPARTMENT OF EDUCATION
BASIC SKILLS IMPROVEMENT POLICY

TEST REVIEW FORM #1
CONTENT MATCH TO STATE READING OBJECTIVES

Test Name _____

Form and Level _____
Copyright _____
Publisher _____
Contact Person _____
Phone Number _____

Directions: Next to each of the state's secondary level basic skills objectives, specify the item number(s) on this test that have been judged to measure the objective.

Reading Objectives	Test Item Number(s) (completed by publisher)	Reviewer's Response	
		Yes	No
1. Identify the meaning of commonly used words within a sentence that does not provide clues to the meaning of the word			
2. Identify the meaning of a word within a sentence that provides clues to the meaning of the word			
3. Identify the meaning of a written phrase, clause, sentence, or paragraph			
4. Demonstrate the ability to follow directions			
5. Identify the main idea, supporting details, and conclusion of a paragraph			
6. Recognize the sequence of events or ideas in a written passage			
7. Identify information on a chart, map, or graph			
8. Draw conclusions implied in a paragraph or passage			
9. Identify cause and effect relationships implied in a paragraph or passage			
10. Predict an outcome implied in a paragraph or passage			
11. Identify a statement as fact or opinion			
12. Identify the writer's purpose in a paragraph or passage written to inform or persuade			
13. Use the parts of a book			
14. Locate information in a variety of sources			

MASSACHUSETTS DEPARTMENT OF EDUCATION
BASIC SKILLS IMPROVEMENT POLICY

TEST REVIEW FORM #2
CONTENT MATCH TO STATE MATHEMATICS OBJECTIVES

Test Name _____

Form and Level _____
Copyright _____
Publisher _____
Contact Person _____
Phone Number _____

Directions: Next to each of the state's secondary level basic skills objectives, specify the item number(s) on this test that have been judged to measure the objective.

Mathematics Objectives	Test Item Number(s) (completed by publisher)	Reviewer's Response	
		Yes	No
1. Recognize number symbols (17, eighteen), whole numbers(34), fractions ($\frac{1}{2}$), decimals (3.75), and powers of 10 (10^2)			
2. Identify odd and even numbers			
3. Put numbers in numerical order			
4. Recognize equivalent fractions			
5. Add, subtract, multiply, and divide whole numbers (4069 + 81 + 123, 254 x 17, 16300 ÷ 100)			
6. Add and subtract mixed numbers			
7. Multiply whole numbers or money by fractions (halves, quarters, thirds)			
8. Add, subtract, multiply, and divide decimal numbers like money			
9. Change a fraction to a decimal ($\frac{1}{4}$ to .25)			
10. Find a percent of a number in situations such as simple interest, discounts, commissions, and taxes			
11. Use ratio and proportion (mixtures, recipes, scale drawings)			
12. Use simple formulas (A=lxw)			
13. Round off numbers to a specified place			
14. Approximate the answer to a computation problem (including discounts and percentages)			

Mathematics Objectives (Cont'd.)	Test Item Number(s) (Completed by Publisher)	Reviewer's Response	
		Yes	No
15. Estimate length, weight/mass, capacity time, temperature, area, and volume			
16. Estimate with money			
17. Choose an appropriate unit of measurement in the U.S. customary system (for example: feet, pounds, and gallons)			
18. Choose an appropriate unit of measurement in the metric system (for example: meters, kilograms, liters)			
19. Choose an appropriate measurement instrument involving both U.S. customary and metric units.			
20. Convert common measurements within the same system			
21. Read a scale drawing			
22. Use a map to compute highway distances			
23. Relate total cost and cost per unit			
24. Compute by using temperature			
25. Compute by using time			
26. Identify right angles and parallel, perpendicular, and intersecting lines like those in a street map			
27. Recognize that an object has the shape of a square, rectangle, triangle, or parallelogram			
28. Identify the radius, diameter, and center of a circle			
29. Recognize that an object has the shape of a cube, cylinder, or sphere			
30. Find the perimeter of a triangle, square, and rectangle			
31. Find the area of a triangle, square, and rectangle			
32. Find the volume of a cube or other rectangular solid			

Mathematics Objectives (Cont'd.)	Test Item Number(s) (Completed by Publisher)	Reviewer's Response	
		Yes	No
33. Read a table			
34. Interpret a bar graph			
35. Interpret a circle graph			
36. Interpret a line graph			
37. Understand probabilities like those used in weather forecasting or lotteries (the chance something will or will not happen)			
38. Find and use averages (mean and median) for a group of numbers			

Appendix B

Bias Review Criteria

BIAS REVIEW CRITERIA:

SECONDARY LEVEL EVALUATION INSTRUMENTS

All evaluation instruments will be reviewed for bias on the basis of these criteria. In reviewing instruments, overall patterns of bias and blatantly offensive content are to be avoided. (All items listed below may not apply in every case.)

1. CHECKLIST FOR RELEVANCY EVALUATION

A. Does the material accurately reflect the students' language experiences?	YES	NO	NA
B. Does the material reflect the students' own experiences?	YES	NO	NA
C. Does the material reflect the different levels (formal, deep, situational, etc.) of the students' culture?	YES	NO	NA
D. Does the material affirm the worth of the group members as individuals?	YES	NO	NA
E. Does the material reflect the students' culture as vital and dynamic: one in which an individual can grow, be nourished and enriched as a unique human being?	YES	NO	NA

2. CHECKLIST FOR AUTHENTICITY EVALUATION

A. Is the material considered authentic by students or other members of the cultural group depicted?	YES	NO	NA
B. Is comparison of cultures seen in a positive light (i.e. does not in any way imply the inferiority of one culture to another)?	YES	NO	NA
C. Does the material reflect authentic cultural values?	YES	NO	NA

3. CHECKLIST FOR LANGUAGE EVALUATION

A. Is the language used fimiliar to the student?	YES	NO	NA
B. Is the material's use of dialects accurate? Are dialects used without confusing them with one another?	YES	NO	NA
C. Is the usage of dialects viewed as an acceptable way of communication (not used in a derogatory or demeaning way)?	YES	NO	NA

NA = Not applicable

4. CHECKLIST FOR CULTURAL EVALUATION

4.1 Bias

A. Are different cultures portrayed without an obvious tendency to debase one culture as opposed to another?	YES	NO	NA
B. Are demeaning adjectives avoided in describing one culture as opposed to another (primitive, backward, ancient, etc.)?	YES	NO	NA
C. Are cultural group members shown in a variety of types of behavior?	YES	NO	NA
D. Is the group's material culture represented in ways that avoid emphasizing stereotyped categorizations of the group?	YES	NO	NA
E. Are group members presented so as to avoid being perceived as homogeneous and interchangeable(save for sex and age differences)?	YES	NO	NA
F. Are group members shown as possessing a variety of "cultural" traits of character or belief?	YES	NO	NA
G. Are cultural views, attitudes, and values shown as things taught, challenged, tested, and sometimes discarded, rather than as "just naturally" shared and accepted by all the group members?	YES	NO	NA
H. Are the cultural group's relations with other groups explained as responses to need, threat, competition, etc., instead of as dictated by some inherent cultural attitude (such as "hostile," "warlike," etc.)?	YES	NO	NA

4.2 Content Relevance

A. Does the material reflect the variety of attitudes, styles, and roles within the culture?	YES	NO	NA
B. Does the material recognize both differences and similarities between cultures?	YES	NO	NA
C. Does it reflect the cultyral ideals, values and goals as well as the material and institutional constructs?	YES	NO	NA

D. Is the group's intellectual and artistic expression represented by its members' achievements in forms shared by other groups as well as those unique to itself? YES NO NA

E. Are the group's cultural traits (whether material, artistic or attitudinal) explained as having evolved through interaction with the environment or with other cultural groups? YES NO NA

F. Does the material emphasize the worth of cultural differences as well as similarities? YES NO NA

G. Is the material accurate in its descriptions and explanations of cultural traits, values, behavior? (Mexicans do not celebrate their independence Day on the Cinco de Mayo.) YES NO NA

H. Is the fact of a particular culturally and linguistically distinct group's existence within United States society adequately presented? (Native Americans in New York City and Puerto Ricans in New Mexico, for example, are rarely noticed.) YES NO NA

5. CHECKLIST FOR POLITICAL BIAS

A. Does the material present an accurate and objective account of the conflicts between the United States and other nations? YES NO NA

B. Does the material present an accurate and objective account of the aims, needs and conditions of political protest within the United States? YES NO NA

C. Does the material present an accurate account of the history of the democratic form of government? YES NO NA

6. CHECKLIST FOR RACIAL BIAS

6.1 Bias

A. Does the material avoid debasing one race as opposed to another? YES NO NA

B. Are consistent portrayals of subservient relationships between racial groups avoided? YES NO NA

C. Are persons of all races shown as members of a variety of social, religious or political groups and economic classes? YES NO NA

D. Are members of one race not depicted as experiencing a narrower range of feelings and emotions than members of another race? YES NO NA

E. Are all members of a racial group not presented as sharing in "typical" attitudes or feelings? YES NO NA

F. Are color labels which are used to indicate value ("black" for evil; "white" for pure, etc.) avoided? YES NO NA

G. Are members of a racial group not stereotyped as to manner of speech, style of dress, preferred diversions, etc.? YES NO NA

H. Are inter-racial contacts not pictured consistently as dangerous or hostile? YES NO NA

I. Are persons of one race not presented as having inherent abilities which are denied to persons of other racial groups (non-whites better suited to brutal labor than whites; non-whites more "in tune with nature," etc.)? YES NO NA

6.2 Content Relevance

A. Does the material reflect the racial diversity of United States society? YES NO NA

B. Is an adequate and objective account of racial discrimination presented? YES NO NA

C. Is an adequate account of cultural, historical and social contributions of all racial groups in United States society presented? YES NO NA

D. Does the material reinforce the value of interethnic contact? YES NO NA

E. Are members of all racial groups presented as equally complex and worthwhile human beings? YES NO NA

F. Is the diversity of individual experience within a racial group adequately represented? YES NO NA

7. CHECKLIST FOR HISTORICAL BIAS

7.1 Bias

A. Does the material avoid suppressing or distorting the presence, contribution and point of view of culturally distinct groups to the United States' history? YES NO NA

B. Are known historical factors not suppressed in order to make the actions of some group or person seem good or bad? YES NO NA

C. Are different peoples and cultural groups presented as people struggling to satisfy their own needs and defend their own values rather than as obstacles in the path of another civilization? YES NO NA

D. Is the North American continent not depicted as essentially unsettled until the advent of white Europeans? YES NO NA

E. Does the material not ignore the cost to certain groups of United States historical development (slavery of Black people, loss of land to Native Americans; and economic subjugation of Hispanos in the Southwest; unfair taxation, exploitation and discrimination against Asian labor, etc.)? YES NO NA

F. Does the material avoid representing the rights, goals and needs of one group of people as less worthy than those of another group in the United States' history? YES NO NA

8. CHECKLIST FOR RELIGIOUS BIAS

8.1 Bias

A. Does the material avoid debasing one religion as opposed to another? YES NO NA

B. Does the material avoid proselytizing for a religion or sect? YES NO NA

C. Does the material avoid slighting references to "primitive" religious beliefs or "superstitions"? YES NO NA

D. Are religious symbols, icons or rituals not discussed in a derogatory or demeaning manner? YES NO NA

8.2 Content Relevance

A. Does the material reflect and encourage equal respect for all religious views? YES NO NA

B. Does the material explore the common elements found in various religions? YES NO NA

9. CHECKLIST FOR SEXUAL BIAS

9.1 Bias

A. Does the material avoid debasing one sex as opposed to the other? YES NO NA

B. Does the material avoid showing females only as background figures in a male-centered story? YES NO NA

C. If females are central figures are they shown in roles other than the stereotype of teacher, mother, nurse, and/or secretary? YES NO NA

D. Are the actions of women as sophisticated and as interesting as those of men? (Are women shown using initiative and imagination rather than being docile and passive?) YES NO NA

E. Are the values of beauty, sweetness, domesticity, motherhood and marriage avoided as depictions of the predominant virtues for females in the stories? YES NO NA

F. Does the material avoid showing male figures experiencing themselves as worthwhile only because they shoulder a female's burden or solve her problems? YES NO NA

9.2 Content Relevance

A. Does the material reflect the conditions and contributions of women in United States society? YES NO NA

B. Does the material encourage both boys and girls, and men and women to see themselves as human beings with an equal right to all benefits and choices provided by society? YES NO NA

C. Does the material present an objective account of women's struggle (both historically and contemporarily) to achieve equality with men? YES NO NA

D. Does the material encourage girls to see themselves as capable of filling any role in society? YES NO NA

E. Does the material present an accurate picture of women in other cultures, including their contributions to the culture and their struggles for equality within it? YES NO NA

10. CHECKLIST OF SOCIAL BIAS

A. Does the material avoid debasing one social group as opposed to another? YES NO NA

B. Does the material avoid showing only the nuclear family with middle class values as a worthwhile model for family structure? YES NO NA

C. Do the material's illustrations avoid depicting food and dress that are common only to middle class families? YES NO NA

D. Does the material avoid depicting persons of certain cultural groups as being poor without any explanation for the fact? YES NO NA

The following sources were used in compiling the above criteria:

Materials from:

The Lau Assistance Center, California

The State Department of Education, California

The Far West Laboratory, California

Issues of Sex Bias and Sex Fairness in Career Interest Measurements
H.E.W., National Institute of Education, 1975.

Appendix C

Guidelines for Reviewing Item Quality

Guidelines for Reviewing Item Quality

1. Is the item stem clearly written for the intended group of students?

2. Is the item stem free of irrelevant material?

3. Is a single problem clearly defined in the item stem?

4. Are the answer choices clearly and concisely written for the intended group of students?

5. Is there a correct answer or a clearly best answer?

6. Are distractors free of clues to the correct answer?

7. Was the correct answer randomly positioned among the available answer choices?

8. Do the item stem and answer choices follow standard rules of punctuation and grammar?

9. Does the stem avoid clues to the correct answer (e.g. grammar, content)?

NOTE: An answer of yes to these questions indicate the item meets standard rules of item writing.

Appendix D

Technical Quality Criteria

MASSACHUSETTS DEPARTMENT OF EDUCATION
BASIC SKILLS IMPROVEMENT POLICY

TEST REVIEW FORM #3
TECHNICAL INFORMATION

Test Name _____

Form and Level _____
Copyright _____
Publisher _____
Contact Person _____
Phone Number _____

Technical Criteria	Possible Points	Publisher's Response		Verification (give manual and page number)	Reviewer's Response	
1. Does the technical information accompanying the instrument (e.g. technical manual) include information regarding the following topics:						
a. Item Review Methods	1	Yes	No		Yes	No
b. Item Analysis	1	Yes	No		Yes	No
c. Average Item Difficulty	1	Yes	No		Yes	No
d. Test Validity	1	Yes	No		Yes	No
e. Test Reliability	1	Yes	No		Yes	No
f. Standard Error of Measurement	1	Yes	No		Yes	No
g. Consistency of Mastery Decisions	1	Yes	No		Yes	No
h. Procedures for screening items for offensive sexual, cultural, racial, and/or ethnic content, and/or stereotyping	1	Yes	No		Yes	No

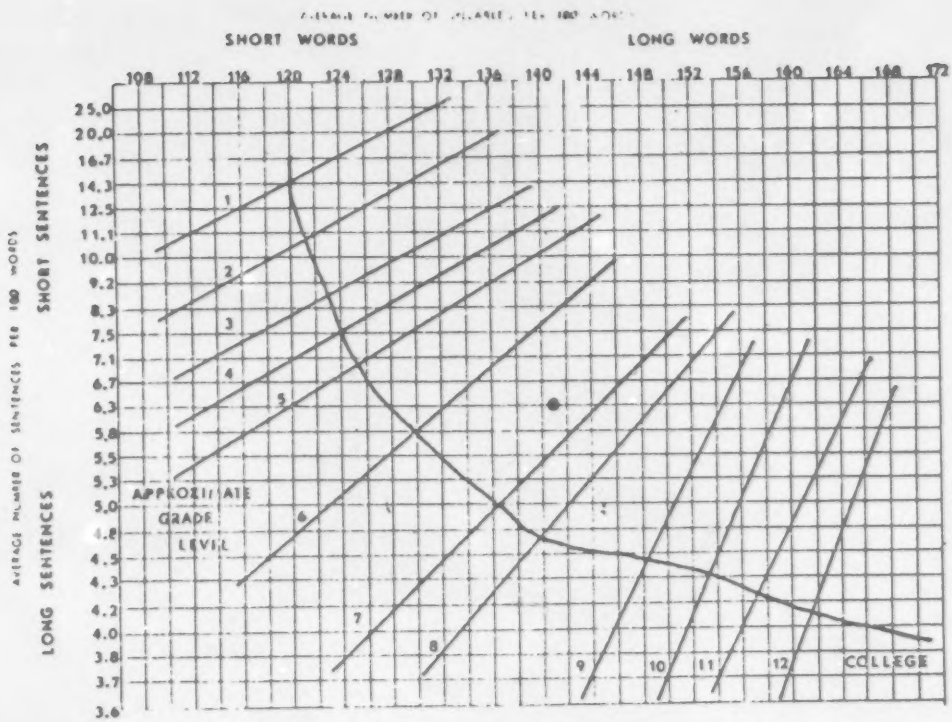
Technical Criteria	Possible Points	Publisher's Response	Verification (give manual and page number)	Reviewer's Response
2. Were item analysis results used to identify defective test items?	3	Yes No		Yes No
3. Do standard indices of internal consistency reliability as reported on the total test score reach or exceed .80?	4 2 1	.90-1.00 .85-.89 .80-.84		.90-1.00 .85-.89 .80-.84
4. Do standard indices of test/ retest reliability reach or exceed .80?	4 2 1	.90-1.00 .85-.89 .80-.84		.90-1.00 .85-.89 .80-.84
5. Were the samples of students used in the field test representative of students in the grades for which this test is intended and representative of important strata within the society (i.e., rural pupils, minority group pupils, pupils in large city schools, etc.)?	2	Yes No		Yes No
6. Are the test administration directions suitable for students in the lowest grade covered by the test?	2	Yes No		Yes No
7. Do the test administration directions provide information to students on time limits?	2	Yes No		Yes No
8. Do the test administration directions tell the student how to handle the problem of guessing?	2	Yes No		Yes No
9. Is the layout of the test booklet convenient for students in the lowest grade covered by the test?	2	Yes No		Yes No

Technical Criteria	Possible Points	Publisher's Response		Verification (give manual and page number)	Reviewer's Response	
10. Is the layout or format of the answer sheet appropriate for students in the lowest grade covered by the test?	2	Yes	No		Yes	No
11. Does the test include practice questions?	2	Yes	No		Yes	No

Appendix E

Readability Formulas

THE FRY READABILITY FORMULA



Directions: Randomly select three 100-word passages from a book or an article. Determine both the average number of words per sentence and the average number of syllables per 100 words.

For example:	*Syllables*	*Sentences*
first 100 words	*124*	*6.6*
second 100 words	*141*	*5.5*
third 100 words	*158*	*6.8*
Average	*141*	*6.3*

On the vertical scale of the above graph find the number you have determined to be the average number of sentences per 100 words. Draw a horizontal line through the graph at this point. Similarly, on the horizontal scale locate the number you have determined to be the average number of syllables per 100 words. Draw a verticle line through the graph at this point. Put a dot at the point your two lines intersect. The grade level line closest to the dot indicates the level of the text - - in the example the level is seventh grade.

Edward Fry, "Readability Formula That Saves Time," Journal of Reading, 11 (April, 1968), pp 513-16; 575-78

SMOG GRADING FORMULA

SMOG Grading is simply a readability formula devised by Dr. G. Harry McLaughlin, Associate Professor of Communications in the School of Journalism, Syracuse University.

1. *Count 10 consecutive sentences near the beginning of the text to be assessed, 10 in the middle and 10 near the end. Count as a sentence any string of words ending with a period, question mark or exclamation point.*

2. *In the 30 selected sentences count every word of three or more syllables. Any string of letters or numerals beginning and ending with a space or punctuation mark should be counted if you can distinguish at least three syllables when you read it aloud in context. If a polysyllabic word is repeated, count each repetition.*

3. *Estimate the square root of the number of polysyllabic words counted. This is done by taking the square root of the nearest perfect square. For example, if the count is 95, the nearest perfect square is 100, which yields a square root of 10. If the count lies roughly between two perfect squares, choose the lower number. For instance, if the count is 110, take the square root of 100 rather than that of 121.*

4. *Add 3 to the approximate square root. This gives the "SMOG Grade" which is the reading grade that a person must have reached if he is to understand fully the text assessed.*

McLaughlin, G. Harry. "SMOG Grading -- a New Readability Formula," Journal of Reading, 12 (May, 1969), pp. 639-645.

Appendix F

Calculation of Standard

Error of Measurement at the Cutoff

STANDARD ERROR OF MEASUREMENT

At the Cutoff

The computational formula for the standard error of measurement for a given score is:

$$\text{Standard Error} = \sqrt{\frac{X\,(N-X)}{N-1}}$$

Where

N = the total number of items on the test

X = the score for which the standard error is desired

This formula can be reduced to the following five steps:

STEP 1

Choose the score for which you want the standard error of measurement. This may be your cutoff score, or a score you are considering as a cutoff score, or simply the score of an individual student. As an example, let us assume we have chosen a cutoff score of 30 on a 100 item test.

STEP 2

Subtract the score chosen in Step 1 from the total number of items on the test.

$$100 - 30 = 70$$

STEP 3

Multiply the number obtained in Step 2 by the score chosen in Step 1.

$$70 \times 30 = 2100$$

STEP 4

Divide the number obtained in Step 3 by the number of items on the test minus one (N-1).

$$\frac{2100}{99} = 21.21$$

STEP 5

Take the square root of the number obtained in Step 4. The result is the standard error associated with the score selected in Step 1.

$$\sqrt{21.21} = 4.60$$

Standard error associated with a score of 30 on a 100 item test = 4.60

Appendix G

Internal Consistency Reliability Coefficient

KR-20 Reliability Coefficient

The computational formula for the KR-20 reliability coefficient is

$$r = \frac{n}{n-1} \quad \frac{V - \Sigma pq}{V}$$

where

n = number of items on the test

V = the variance of scores on the test defined as

$$\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N-1}$$

where N = number of test scores

ΣX = sum of the test scores

ΣX^2 = sum of the squared test scores

pq = the proportion of correct responses for a given item times the proportion of incorrect responses for that item

Σpq = the sum of the above products for all items

The above formula, however, can easily be reduced to a series of simple computational steps. For those unaccustomed to using statistical formulas, we recommend following the steps as they are outlined below.

Step 1

Create a table of students' raw scores on the test (i.e., number of items answered correctly), such as the one below.

Table 1

Student	Score
Everett	5
Chris	4
Don	3
Susan	1

Step 2

Create a second table, such as the one below, which shows the proportion of students that answered each item correctly and the proportion that answered each item incorrectly.

Table 2

Item	Proportion of Students Who Answered Correctly	Proportion of Students Who Answered Incorrectly
1	1.00	0.00
2	.75	.25
3	.75	.25
4	.50	.50
5	.25	.75

Step 3

Add all the scores in Table 1

$$5 + 4 + 3 + 1 = 13$$

Step 4

Square the value obtained in Step 3 and divide by the number of students in Table 1.

$$13^2 - 4 = 42.25$$

Step 5

Square each score in Table 1 and add the squared values.

$$5^2 + 4^2 + 3^2 + 1^2 = 51$$

Step 6

Subtract the results of Step 4 from the results of Step 5.

$$51 - 42.25 = 8.75$$

Step 7

Divide the results of Step 6 by the number of students in Table 1 minus 1.

$$8.75 \div 3 = 2.91$$

Step 8

In Table 2 multiply the proportion who answered each item correctly by the proportion who answered it incorrectly and sum the products.

$$(1 \times 0) + (.75 \times .25) + (.75 \times .25) + (.50 \times .50) + (.25 \times .75) = .81$$

Step 9

Subtract the result of Step 8 from the result of Step 7.

$$2.91 - .81 = 2.1$$

Step 10

Divide the result of Step 9 by the result of Step 7.

$$2.1 \div 2.91 = .72$$

Step 11

Divide the number of items on the test by one less than the number of items on the test.

$$5 \div 4 = 1.25$$

Step 12

Multiply the result of Step 10 by the result of Step 11. The result is the internal consistency coefficient.

$$r = .72 \times 1.25 = .90$$

Appendix H

Calculation of the Pearson Product-Moment Correlation

Pearson Product-Moment Correlation

The computational formula for the Pearson product-moment correlation is

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

where

N = number of pairs of test scores

ΣXY = sum of the products of the paired scores

ΣX = sum of the scores from the first test

ΣY = sum of the scores from the second test

ΣX^2 = sum of the squared scores from the first test

ΣY^2 = sum of the squared scores from the second test.

The above formula, however, can easily be reduced to a series of simple computational steps. For those unaccustomed to using formulas, the following step-by-step approach will yield the same results.

Step 1

Arrange the scores in a table by pairs as follows. Each student must have two scores.

- For test/retest reliability, the first score will be the first administration of the test and the second score will be the second administration of the test.
- For interrater reliability when scoring writing or speaking samples, the first score will be the score assigned by the first rater and the second score will be the score assigned by the second rater.
- For interform reliability, the first score will be the score on the first form and the second score will be the score on the second form.

	First Score	Second Score
Student #1	20	20
Student #2	44	41
Student #3	32	32

Step 2

Multiply the two scores in each pair; then add the products.

(20 X 20) + (44 X 41) + (32 X 32) = 3228

Step 3

Multiply the number obtained in Step 2 by the number of students.

3228 X 3 = 9684

Step 4

Square each score in the first column, and add the squared values.

$20^2 + 44^2 + 32^2 = 3360$

Step 5

Multiply the sum of Step 4 by the number of students.

3360 X 3 = 10080

Step 6

Add all the scores in the first column.

20 + 44 + 32 = 96

Step 7

Square the value obtained in Step 6.

$96^2 = 9216$

Step 8

Square each score in the second column, and add the squared values.

$20^2 + 41^2 + 32^2 = 3105$

Step 9

Multiply the result of Step 8 by the number of students.

3105 X 3 = 9315

Step 10

Add all the scores in the second column.

$20 + 41 + 32 = 93$

Step 11

Square the value obtained in Step 10.

$93^2 = 8649$

Step 12

Multiply the final value of Step 6 by that of Step 10.

$96 \times 93 = 8928$

Step 13

Subtract the value obtained in Step 12 from that obtained in Step 3. Be careful to note whether the sign is positive or negative.

$9684 - 8928 = 756$

Step 14

Subtract the value of Step 7 from that of Step 5.

$10080 - 9216 = 864$

Step 15

Subtract the value of Step 11 from that of Step 9.

$9315 - 8649 = 666$

Step 16

Multiply the result of Step 14 by the result of Step 15.

864 X 666 = 575424

Step 17

Take the square root of the result of Step 16.

$\sqrt{575424} = 758$

Step 18

Divide the value of Step 13 by that of Step 17. The result is the Pearson product-moment correlation which summarizes the relationship between the two sets of scores.

r = 756/758 = .99

Adapted from:

Bruning, J, and Kintz, B. Computational Handbook of Statistics. Glenview, Ill.: Scott, Foresman and Company, 1977.

Appendix I

Sample Technical Manual

TECHNICAL MANUAL
SEAPORT, MASSACHUSETTS
EIGHTH GRADE BASIC SKILLS TEST IN MATHEMATICS

I. OBJECTIVES

For its Eighth Grade Basic Skills Test in Mathematics, Seaport has adopted the basic skills objectives in reading and mathematics from Section 40.04 of the Basic Skills Improvement Policy and Regulations. Item specifications were written for the word problem objective by an Item Committee comprised of ten seventh and eighth grade teachers. Other items were written directly from objectives.

II. ITEM REVIEW

Quality

All items were written by the Item Committee. These items were then reviewed and revised by a separate Item Review Committee using the item quality guidelines supplied by the Department of Education. The Item Review Committee was comprised of two junior high school administrators, two junior high school teachers, and the central office administrator. In the end, all items selected for final use were judged to meet the above criteria for good items.

Content Validity

As a check for content validity, the Item Review Committee matched each item with an objective from Section 40.04 of the Basic Skills Improvement Policy. Results of this matching exercise were reviewed by the Item Committee. Items which could not be matched or were incorrectly matched were revised and resubmitted to the Item Review Committee.

Bias

Each item was reviewed by the Item Review Committee for relevance, authenticity, and language, and for cultural, political, racial, historical, religious, sexual, and social bias. Each member of the committee reviewed for bias in two of the above categories. This review employed the Bias Review Criteria contained in Appendix C of the Basic Skills Improvement Policy Implementation Guide #1.

Pilot Testing

Each teacher on the Item Committee and the Item Review Committee administered a portion of items to the students in his/her homeroom class. Feedback on the items and test directions was solicited from the students via class discussion. Items which students found confusing or deceptive were subsequently revised by the Item Committee.

III. FIELD TESTING

Sampling Plan

All eighth grade students in the district who attended school on April 15, 1981, took part in the field test. Students were tested during an extended homeroom period in the morning. Two sets of items were printed. Each set contained 75 items. Teachers were instructed to alternate sets as they passed out the test, so that distribution of the sets in the classroom would be random. In all, 223 students took part in the field test.

Item Analysis

Item statistics were obtained as part of the scoring service. It was decided to drop or revise items for which more than 90% of the students chose the wrong answer, and items on which more than 20% of the students chose the same wrong answer. No such items were found in either test form.

IV. TEST FORMS

Format

Test forms were constructed so that each objective would be measured by at least one item. Items were assigned to forms based on the item difficulties obtained during field testing. An attempt was made to equalize the difficulty of the two forms based on these preliminary values.

There were 45 items in each form of the mathematics test. Test booklets were modeled after Level 18 (8th grade) of the California Achievement Test. Each form was laid out to allow maximum spacing between items without continuing items onto a second page. Test booklets were printed by a professional copying service.

Answer sheets were purchased from Danville State College so that the results could be machine scored using their equipment. These answer sheets were especially designed for use with elementary level achievement tests.

Directions

Directions were written by the Item Committee, reviewed by the Item Review Committee, and pilot tested along with items to obtain

student feedback. Directions were also pilot tested individually on a few sixth grade readers to insure that they would be readable to eighth grade students. Directions contained time limits for the test, and instructed students to guess when they were in doubt about the correct answer.

Sample Items

Both forms of the test began with three sample items which were completed under the direction of the teacher before testing began.

Bias Review

Once test forms had been constructed, they were reviewed for bias by the Item Review Committee. Although individual items had already been screened, this second review insured that no set of items in the test would violate the bias review criteria outlined in Implementation Guide #1.

TEST STATISTICS

The following test statistics were obtained during the first administration of the completed test on October 2, 1981.

Difficulty

The average item difficulty for Form A was .81. The average item difficulty for Form B was .78.

Standard Error of Measurement at the Cutoff

Because Seaport has yet to establish standards on this test, the standard error of measurement was calculated at three potential cutoff scores. The average of these three was 3.0 on Form A and 3.1 on Form B.

Internal Consistency Reliability

The KR-20 internal consistency reliability coefficient on Form A was .91. On Form B, it was .89.

Test/Retest Reliability

The same form of the test was readministered to a random sample of sixty students on October 16, two weeks after the initial test. Half of them had initially taken Form A and the other half Form B.

Test/retest reliability coefficients for Forms A and B were .96 and .93, respectively.

Parallel Form Reliability

On October 16, two weeks after the initial test, thirty students were retested with the form they had not taken on October 2. Half of them had initially taken Form A and the other half Form B.

The interform reliability coefficient was .90.

Consistency of Mastery Decisions

Because Seaport has yet to establish standards for this test, the consistency of mastery decisons was examined at three "potential" cutoff scores (15, 20, and 25 items correct). Students in the test-retest reliability sample were classified as masters or non-masters on both administrations of the test using each of these three possible cutoff scores. On the average, they were consistently classified 91.5% of the time.